Audio-Perceptual Evaluation of Portuguese Voice Disorders—An Inter- and Intrajudge Reliability Study

*,†,‡Susana Vaz Freitas, †Pedro Melo Pestana, §Vítor Almeida, and §Aníbal Ferreira, *†‡§*Porto, Portugal*

Summary: Objectives/Hypothesis. The aim of this article was to describe the results of an audio-perceptual evaluation carried out by 10 judges, on a database comprising 90 voice recordings plus 10 samples repetition, with the purpose of characterizing the intra- and interrater reliability.

Study Design. Exploratory, transversal.

Methods. The classification of the GRBAS parameters was obtained for each one of the 10 experts, concerning the 90 voice samples. The intraclass correlation coefficient determined the interrater reliability. For the 10 repeated voices, the intrarater reliability was assessed by means of a dispersion analysis.

Results. The average judges' classification for each of the GRBAS parameters differs ($P < 0.05$). The values of the correlations, with confidence intervals of 95%, between the average scores for all components of the GRBAS scale lie, in general, between 0.838 and 0.966. The first three parameters of the scale (G, R, and B) have the higher interrater reliability. Differences were statistically significant ($P < 0.05$) for experts 1, 6, 9, and 10, which means a poor intrarater reliability for 40% of the judges.

Conclusions. All the experts had similar evaluation criteria for the assessment of the five parameters of the GRBAS scale (the values of the confidence intervals at 95% of the experts average ratings of the GRB were above 0.8). However, its quantification is not statistically similar. Asthenia and Strain have lower reliability. Most experts do not reveal statistically significant differences between the values assigned to the GRB parameters ($P > 0.05$).

Key Words: Audio-perceptual–Voice–Assessment–Reliability.

INTRODUCTION

The audio-perceptual evaluation assumes that the voice professional evaluates a vocal sample produced by the speaker who refers (or not) complaints about voice use. This evaluation method is widely used on an everyday clinic environment for many reasons, such as the fact that the vocal quality is perceptible in its nature, and, thus, its features assume an intuitive value and have the possibility of being shared among the listeners.[1,2] Generally, the vocal exercises asked to a patient include: sustained vowels and connected speech, which are later recorded in audio and/or video (preferably) format for a database construction and demonstrative support (to the patient) and for comparison with future evaluations.

According to some authors, the perceptual evaluation of pathologic voices is a core component of the process of dysphonia characterization[3] and, it is, by far, most commonly used to describe the voice in a clinical setting[4]: essentially due to the fact that it is fast and efficient and it implies few material resources (ie, it is low cost).[5,6] Still, there are some problems in using the audio-perceptual evaluation of vocal quality,[1] namely:

1. The low intra- and interjudges consistency;
2. It does not provide objective measures;
3. The nonexistence of a universal scale of perceptual evaluation.

The literature review shows that the accuracy of perceptual evaluation relies on several factors.[1,5,7–10] Kreiman and Gerratt[7] suggest that the use of perceptual scales may sustain errors and variability because (1) scales used in clinical and research settings are, sometimes, not the proper ones to measure the voice quality attributes; (2) judges do not always agree on the parameters that are being evaluated; (3) judges are not always able to find one single dimension of the scale in a complex sound stimulus; (4) judges tend to exhibit low consistency in their classifications, intra- and interevaluators.

The recording and analysis of vocal samples may be implemented in a formal way (using protocol scales) or in an informal way (by analyzing the patients' voice features, with focus on the different intervening voice production systems—breathing, phonation, articulation, and resonance). This is an integrated process that consists, mainly, in hearing and describing a given voice, referring to its features in general terms or with focus on specific parameters, which might be associated to psychoacoustic and pathologic features.[8]

Literature accounts for the use of many scales since the 80s. According to Hammarberg[4] and Cummings,[3] the most used and broadly known scale is the GRBAS, by Hirano.[9] This scale was developed and implemented in 1969 by the Committee for Tests of Phonatory Functions of the Japan Society of Logopedics and Phoniatrics, based on the research studies carried out by Isshiki et al.[10] The acronym GRBAS is composed by each one of the five graphemes: G, Grade of Dysphonia; R, Roughness; B, Breathiness; A, Asthenia; and S, Strain.[9] It is a compact and easy to use scale, very effective for vocal screening, which evaluates glottic source during the production of sustained vowels (/a/ or /ɛ/), reading, or connected speech. The evaluated

Accepted for publication August 2, 2013.

From the *Neuroscience Department, Speech Pathology Unit of Otolaryngology Service, Centro Hospitalar do Porto, Porto, Portugal; †Speech Therapy Department, Faculty of Health Sciences, Universidade Fernando Pessoa, Porto, Portugal; ‡Biomedical Engineering Department, Faculty of Engineering, Universidade do Porto, Porto, Portugal; and the §Electrical and Computer Engineering Department, Faculty of Engineering, Universidade do Porto, Porto, Portugal

Address correspondence and reprint requests to Susana Vaz Freitas, Neuroscience Department, Centro Hospitalar do Porto, Serviço de Otorrinolaringologia, Largo Prof. Abel Salazar, 4099-001 Porto, Portugal. E-mail: svazfreitas@gmail.com

Journal of Voice, Vol. 28, No. 2, pp. 210-215
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2013.08.001

parameters are classified in a four-point scale that discriminates severity levels: 0 = normal or absence of perturbations; 1 = slight or discreet perturbations; 2 = moderate or evident perturbations; and 3 = severe/extreme perturbations.

The Vocal Profile Analysis Scheme is widely used by speech therapists in the UK, based on the work of Laver et al.[11] This protocol is an approach to the measurement of voice quality, supported by a theoretical approach featuring voice according to its physiology.[7] It enables the description of laryngeal and supraglottic features (vocal tract) with 31 parameters divided into three subcategories. The final score on this scale is presented in six degrees.

Over the last three decades, the Department of Speech and Language Pathology of the Huddinge Hospital (Sweden) has developed and improved the Stockholm Voice Evaluation Approach (SVEA).[4] This scale was based on the analysis of correlations between 28 variables (based on 50 perceptive terms used by the clinicians), resulting in 13 parameters proposed for the qualitative voice assessment. It is quantified into five levels (where 0 = normal and 4 = very severe).

Consensus Auditory Perceptual Evaluation of Voice is an audio-perceptual rating scale that ranks six vocal parameters (Global Severity, Roughness, Breathiness, Strain, Pitch, and Loudness) with resource to a 100-mm visual analog scale (complemented by other descriptors: consistency/inconsistency), as well as two additional voice data, such as classification of resonance or tremor. It was developed in 2002, after a conference of the American Speech-Language-Hearing Association. It is based on sustained vowels (/a/ and /i/, for 3–5 seconds), reading preset phrases and spontaneous conversation. Instructions for its use and quotation are available online, at the American Speech-Language-Hearing Association's Division 3 for Voice and Voice Disorders.[12,13]

The use of these scales has been somewhat criticized, mainly because they do not consider the classification of some characteristics of the voice.[7] Some studies identified low reliability of the Asthenia and Strain parameters.[14–17] However, this method correlates moderately with other forms of vocal classification, mainly the questionnaires that measure the impact of the dysphonia in the quality of life.[18–22]

The aim of this article was to describe the results of the evaluation carried out by 10 experts in audio-perceptual evaluation of a database consisting of 90 voices and 10 of these samples randomly selected repetition. The inter- and intrajudge reliability are statistically analyzed and the results are discussed.

METHODS

Ten Speech-Language Pathologists (three Brazilian and seven Portuguese) specialized in voice disorders assessment and intervention were recruited as judges to listen, analyze, and classify perceptually 100 voice samples. The contact was established by the first author and confirmed by e-mail and/or in person.

The selection was based on the following criteria: years of professional experience (>6 years)[23] in a hospital setting or with professional voice and/or with academic responsibilities in the area of Voice, with knowledge and use of GRBAS. The gender distribution is even (50%). The average number of years of experience is 12.6 years. Inclusion criteria were the absence of a history of language or speech disorders, as well as history of hearing loss or hearing disorders at the time of completion of the evaluation.

Ninety voices were selected from a database with the consent of the Hospital's Ethics Committee. Of these, 20 were normal voices and 70 had some degree of disturbance (in different degrees of severity, from mild to severe).

The gender distribution was performed as follows: 28% of male voices (n = 25) and 72% of female voices (n = 65). This distribution corresponds to the representation of each of these genders in the database. Concerning the disturbed voices, gender representation is 34% male and 66% female speakers. All individuals were adults, aged over 18 years old.

The order of presentation of the vocal stimuli was determined randomly, so as to avoid the effects of familiarity. To these 90 original voices, 10 were added again. These 10 productions were randomly selected and emerged from the disturbed and normal voice pools.

Vocal stimuli recordings followed the same sampling protocol, which implied: a sampling frequency of 44 100 Hz and a resolution of 16 bits per sample, using a Philips SBC ME 400 (Philips, Amsterdam, The Netherlands) desktop microphone, unidirectional (cardioid), a room with a noise level lower than 40 dB SPL, although not acoustically treated.

The distance from the microphone to the mouth was fixed at 10 cm and the patient was asked to produce a sustained vowel [α], after illustration by the voice specialist, at a comfortable level of pitch and loudness,[5,24,25] during at least 5 seconds,[5,25,26] in two trials. The last one, with the speaker standing up, was used for this study. These recordings are part of the routine evaluation protocol of voice hospital appointments by the first author and were collected using *Dr. Speech* software, Version 4.0 (Tiger Electronics, USA). The portion of the signal from the 2nd second until the end was segmented and analyzed[24–26] and it was considered to be the most stable signal region for this research.

The audio-perceptual evaluation of the 100 voices was presented on a Web page, randomly ordered, heard, and rated according to the same interactive sequence by a panel of 10 judges. The Web page presents a form for every voice to be assessed, with a pre-allocation of the audio file to be heard and a set of five adjustable interactive cursors—one for each parameter of the GRBAS scale. The classification is made in a visual analog scale (VAS), represented by a 100-mm long ruler (the more to the right, the more disturbed the voice quality is judged), so that the experience of evaluating each parameter is closer to an analogic representation. The VAS was chosen due to the results put forward by Yiu et al[27] and, recently, Karnell et al,[21] who stated that the assessment with the VAS scale shows more consistent results than with the Equal Appearing Interval (EAI) scale. Each stimulus could be heard repeatedly (which was also saved), according to the needs of each judge in defining the classification of the five audio-perceptual parameters. An illustration screenshot of this web interface is

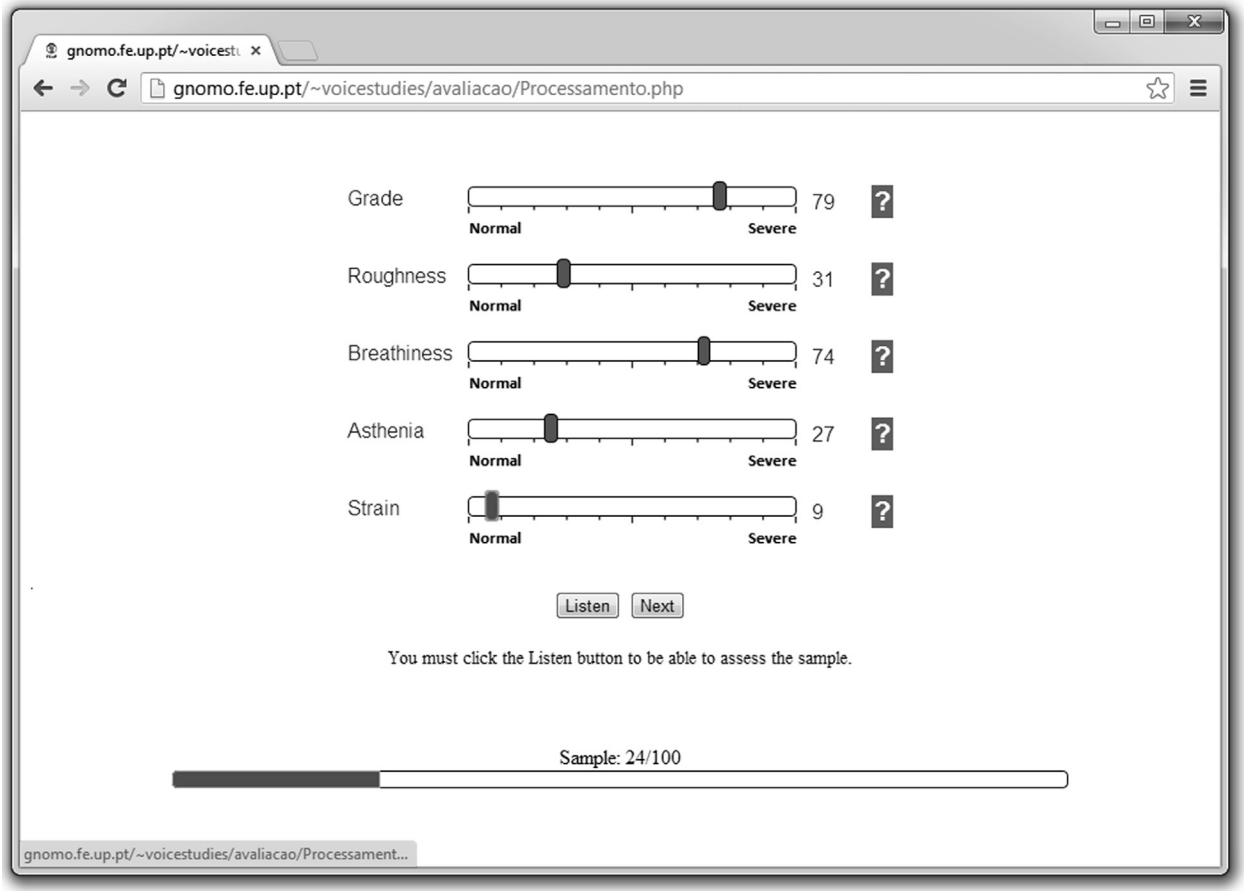

FIGURE 1. Example of the Web interface used to collect audio-perceptual evaluation.

presented in Figure 1. When all fields had been classified, the evaluator could go back and correct any of the answers or proceed down the page.

The statistical analysis was carried out using the computer program *Statistical Package for the Social Sciences*—IBM SPSS for Windows, Version 19.0 (IBM, Chicago, IL).

Data analysis was performed in two stages. Initially, depending on the nature of the studied variables, descriptive statistics were calculated—central tendency (mean and median), dispersion (standard deviation), and also the extreme values (minimum and maximum).

In a second phase, after evaluating asymmetry and kurtosis (through their coefficients), and normal distributions (by applying the nonparametric Kolmogorov-Smirnov test, with Lilliefors correction), the assumptions of normality were checked and parametric tests were applied; otherwise, we resorted to the use of nonparametric ones. All tests were applied with a confidence level of 95%, except where explicitly noted.

RESULTS

Descriptive analysis of audio-perceptual evaluation

Table 1 depicts the main descriptive statistics of GRBAS, according to the 10 recruited judges, for the 90 used voice samples. On average, the highest values for each one of the scale parameters were found in different judges.

The high dispersion observed for all studied parameters (standard deviation between 12.00 and 35.00) should be emphasized. Results allow observing that the parameter with the highest ratings was G (Grade) and that the one in which the scores tended to be lower was A (Asthenia).

Analysis of interjudge reliability (intraclass correlation coefficient)

Given that 10 experts were recruited to evaluate the GRBAS components for 90 voices, the intraclass correlation coefficient (ICC) was used to assess the degree of agreement, thus making it possible to determine if all judges interpreted the audio perceptual parameters to classify the GRBAS identically, which would ensure its consistency (Table 2).

It was concluded that the classification in central tendency given by the 10 experts to each component of the GRBAS differed in a statistically significant way (P values <0.05). The values of the confidence intervals at 95% of the correlations between the average scores given by experts for all GRBAS components lie, in general, between 0.838 and 0.966—values greater than 0.8—therefore, there is a good consistency of the parameters of the used scale. It should be noted that the first three parameters of the scale (G, R, and B) have a greater consistency than the remainder (A and S). Thus, it is safe to say that although the 10 judges manifested the same evaluation criteria for the five GRBAS

TABLE 1.
Descriptive Statistics of the GRBAS Scale According to the Judges

	Scale (100-mm VAS)									
	G (Grade)		R (Roughness)		B (Breathiness)		A (Asthenia)		S (Strain)	
Judges	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	48.19	25.00	33.63	29.00	36.32	25.00	9.98	18.00	45.51	27.00
2	28.40	30.00	13.61	26.00	5.61	15.00	13.56	24.00	13.23	23.00
3	29.46	35.00	14.82	28.00	17.76	25.00	9.60	18.00	12.03	26.00
4	36.59	26.00	34.99	26.00	**51.94**	27.00	21.34	26.00	**51.11**	27.00
5	**56.27**	18.00	37.51	20.00	37.12	27.00	**30.00**	28.00	29.26	30.00
6	39.63	31.00	16.08	24.00	20.13	26.00	17.13	25.00	17.57	22.00
7	47.52	26.00	**42.74**	25.00	43.66	27.00	28.00	31.00	16.43	25.00
8	51.73	25.00	26.92	29.00	40.74	29.00	13.63	21.00	14.73	26.00
9	53.21	23.00	22.01	26.00	40.26	30.00	9.17	18.00	10.71	22.00
10	50.48	17.00	32.87	24.00	32.00	22.00	4.96	12.00	23.08	20.00

Abbreviation: SD, standard deviation.

parameters under analysis; they do not make their quantification in a statistically similar way.

Analysis of the difference between the "original" and "repetition" voices, according to the judges (intrajudge reliability)

The analysis of the classification of each GRBAS parameters, made by the 10 audio-perceptual judges, to the "original" and "repetition" voices (corresponding to 10 voices) of the total sample, reveals discrepancies and uniformity of the judges' ratings. This led to the exclusion of voices 1, 5, and 6 which possess greater assessment variability. It can be noticed that the lowest ICC was obtained for the parameters A and S. The argument is that those particular voice records do not elicit consensus.

Subsequently, a study of the audio-perceptual ratings was carried out, by focusing even more attention on the assessment of the remaining seven voices by the 10 judges. Table 3 presents the mean values and standard deviations of the GRB components for original (G1, R1, and B1) and repetition voices (G2, R2, and B2), 7% of the total sample, according to the 10 judges.

Results show that differences were statistically significant ($P < 0.05$) for judges 1, 6, and 10—the ones that rated the same voice in a significantly different way, on both moments ($P < 0.05$)—which translates into a poor intrajudge reliability. Concerning judge 1, the ratings assigned to components G ($P = 0.028$) and B ($P = 0.043$) were significantly higher in repetitions than in original voices. Similarly, judge 6 rated components G ($P = 0.047$) and R ($P = 0.009$) in a significantly higher way for repetitions than for original. Judge 10 ranked the B parameter ($P = 0.005$) in a significantly higher way for repetitions than for original voices.

For the remaining judges (2, 3, 4, 5, 7, 8, and 9), there were no statistically significant differences between the values assigned to GRB parameters, for the original and repetition voices ($P > 0.05$).

DISCUSSION

The audio-perceptual assessments made by experts have been characterized with statistical descriptive measures, which revealed a discrepancy between the parameters of GRBAS' classifications. Kreiman and Gerratt[7] reinforce this audio-perceptual assessment difficulty in their publications. In particular, the mean classification for B (Breathiness) and S (Strain) was higher for judge 4. Judge 5 classified G (Grade) and A (Asthenia) with an average greater disturbance. Judge 7 assigns the

TABLE 2.
Descriptive Statistics and ICC for the GRBAS' Parameters (100-mm VAS), According to the Judges

Scale	No. of Items (Judges)	Mean (±SD)	Minimum	Maximum	ICC (CI 95%)	*P**
G	10	44.1 (±10.0)	28.4	56.3	0.952 (0.936; 0.966)	<0.05
R	10	27.5 (±10.4)	13.6	42.7	0.938 (0.916; 0.955)	<0.05
B	10	32.6 (±14.0)	5.6	51.9	0.938 (0.924; 0.959)	<0.05
A	10	15.7 (±8.3)	5.0	30.0	0.887 (0.849; 0.919)	<0.05
S	10	23.4 (±14.3)	10.7	51.1	0.879 (0.838; 0.913)	<0.05

Abbreviation: ICC, intraclass correlation coeficient.
* Results according to nonparametric Kruskal-Wallis test, at 95% of confidence intervals.

TABLE 3.
Descriptive Statistics and ICC for the GRB's parameters (100-mm VAS), According to the Judges

GRBAS	Judges																			
	1		2		3		4		5		6		7		8		9		10	
Scale	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
G*	$P < 0.05$		ns		ns		ns		ns		$P < 0.05$		ns		ns		ns		ns	
G1	39.7	32.8	45.7	30.5	40.4	18.8	35.3	29.7	19.6	17.8	45.1	20.9	20.0	26.5	28.9	30.3	25.3	29.9	53.0	31.1
G2	46.0	25.4	29.9	25.3	39.1	13.4	36.0	29.6	18.6	18.6	49.3	23.1	24.3	21.5	28.6	23.6	16.9	29.2	33.4	23.5
R*	ns		ns		ns		ns		ns		$P < 0.05$		ns		ns		ns		ns	
R1	15.4	20.0	17.9	18.2	24.0	23.4	23.6	28.3	14.3	10.3	22.4	17.4	8.6	15.7	6.9	9.4	5.3	9.1	42.1	26.0
R2	20.0	22.0	19.3	22.5	24.1	21.6	21.7	25.2	18.3	19.1	31.9	19.3	10.0	15.3	15.4	17.4	5.0	8.9	42.4	24.7
B*	$P < 0.05$		ns		ns		ns		ns		ns		ns		ns		ns		$P < 0.05$	
S1	23.4	33.0	29.4	38.4	22.9	23.7	27.0	26.7	32.9	33.2	30.4	26.7	0.0	0.0	14.4	20.4	21.6	24.4	5.0	2.7
S2	26.9	31.5	28.0	25.9	27.4	17.0	20.9	19.2	29.7	29.0	34.0	33.1	0.0	0.0	5.7	10.4	14.0	29.2	37.1	23.0

Abbreviations: M, mean; ns, non-significant; SD, standard deviation.
* Results according to the nonparametric Wilcoxon test for two paired samples, at 95% of confidence intervals.

highest ranking to Roughness, and this fact confirms the approach advocated by Moers et al,[28] who consider that this measure is a supraclass that brings together multiple audio-perceptual parameters. Sofranko and Prosek[23] compared three groups with different experiences in audio-perceptual assessment and concluded that singing teachers (in contrast with speech-language pathologists) tend to perform aesthetic judgments that are more focused on their academic/professional experience. The lowest averages were found for judges 3 and 6, and it has been realized that they exhibit comparable clinical experience, mainly with professional voice users.

It should be noted that the analysis of individual GRBAS parameters shows that the most valued is G (Grade). This conclusion was also attained by the studies of Eskenazi et al,[29] Feijoo and Hernandéz,[30] Dejonckere et al,[31] De Bodt et al,[15] Wolfe and Martin,[32] Yu et al,[33] Heman-Ackah et al,[24] Shrivastav et al,[34] and Choi et al.[35] The A (Asthenia) parameter was less evident. It reinforces the audio-perceptual assessment difficulty for this aspect, which was also found by Dedivitis et al,[36] and this is, probably, the reason why it is not part of some of the audio-perceptual protocols (German RBH evaluation scheme[37]; GRB of the European Laryngological Society[16]).

The ICC allows us to conclude that those judges assume a tendency to classify each audio-perceptual parameter similarly. However, the numerical value was assigned in a different statistically significant way ($P < 0.05$). It should be noticed that the values of the confidence intervals at 95% of the correlations between the average ratings of the experts (ICC) are above 0.9 for parameters G (Grade), R (Roughness), and B (Breathiness). A (Asthenia) and S (Strain) parameters have a lower reliability. These conclusions are confirmed by Dejonckere et al,[14] De Bodt et al,[15] Oates,[1] and Moers et al.[28] This reinforces the trend toward greater assessment accuracy for parameters G (Grade) (as in the studies by Choi et al[35] and Moers et al[28]), R (Roughness) (such as in Wolfe and Steinfatt[38]; Wolfe and Martin[32]; Halberstam[39]), and B (Breathiness).[40] Parameters A (Asthenia) and S (Strain) are excluded from some of the rating scales, due to the reduced interrater reliability, as referred above.

Considering the results of the ICC, an analysis of the difference between the classifications obtained for original and repetition voices was carried out because there was evidence that the assessments of the voices were variable, depending on their characteristics. Thus, the analysis of the 10 repeated voice samples revealed a greater consensus among judges' for a total of seven voices, which were then analyzed for dispersion. Finally, we concluded that four judges' made statistically significant different assessments ($P < 0.05$) of the original and repetitions.

It is noticed that the excluded judges were mostly less consistent in the classification of G (Grade) or B (Breathiness) parameters. These factors are considered, in the literature,[41,42] of relatively easy audio-perceptual assessment, which we could not conclude in our study. This may be attributed to the reduced experience of the judges and, perhaps, also to the privileged contact with cases of professional voice users, in which the existence of disturbance does not always happen. It should also be noted that two of these experts are male. As stated by Moon et al,[43] the classification of a particular voice can be influenced by the gender of the speaker.

No known published research analyzed the influence of gender on the evaluator's classification of audio-perceptual parameters.

CONCLUSION

The final considerations about vocal quality assessment reinforce the need for this to be a multidimensional task.

It should be noted that the average ratings of the experts (ICC) who collaborated in this study had values of confidence intervals at 95% of the correlations above 0.9 for parameters G (Grade), R (Roughness), and B (Breathiness). Parameters A (Asthenia) and S (Strain) had a lower consistency. The judges who rated the 10 repeated samples were, mostly (60%), consistent in both evaluations.

It may be concluded that there is still a wide field of research in this area, which must be sustained in a larger group of experts' classifications, and also in larger voice pools. Studies

aiming at confirming these results and the development of clinical useful tools are pertinent. They should be able to circumvent the drawbacks of existing methods, providing researchers, health professionals, and the speakers themselves with more accurate data obtained expeditiously.

REFERENCES

1. Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatr Logop*. 2009;61:49–56.
2. Shrivastav R. Evaluating voice quality. In: Ma E, Yiu E, eds. *Handbook of Voice Assessments*. San Diego: Plural Pub; 2011.
3. Cummings L. *Clinical Linguistics*. Edinburgh: Edinburgh University Press; 2008.
4. Hammarberg B. Voice research and clinical needs. *Folia Phoniatr Logop*. 2000;52:93–102.
5. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice*. 2009;23:341–352.
6. Orlikoff R, Dejonckere P, Dembowksi J, et al. The perceived role of voice perception in clinical practice. *Phonoscope*. 1999;2:89–108.
7. Kreiman J, Gerratt B. Perceptual assessment of voice quality: past, present and future. In: Ma E, Yiu E, eds. *Handbook of Voice Assessments*. San Diego: Plural Pub; 2011.
8. Carding P, Carlson E, Epstein R, Mathieson L, Shewell C. Formal perceptual evaluation of voice quality in the United Kingdom. *Logoped Phoniatr Vocol*. 2000;25:133–138.
9. Hirano M. *Clinical Examination of Voice*. New York: Springer-Verlag; 1981.
10. Isshiki N, Okamura H, Tanabe M, Morimoto M. Differential diagnosis of hoarseness. *Folia Phoniatr Logop*. 1969;21:9–19.
11. Laver J. *The Phonetic Description of Voice Quality*. Cambridge, UK: Cambridge University Press; 1980.
12. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
13. Chan KM. Auditory-perceptual voice evaluation: a practical approach. In: Ma E, Yiu E, eds. *Handbook of Voice Assessments*. San Diego, CA: Singular Publishing Inc.; 2011.
14. Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr (Basel)*. 1993;45:76–83.
15. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80.
16. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
17. Gould J, Waugh J, Carding P, Drinnan M. A new voice rating tool for clinical practice. *J Voice*. 2012;26:e163–e170.
18. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice*. 1999;13:557–569.
19. Hogikyan ND, Wodchis WP, Terrell JE, Bradford CR, Esclamado RM. Voice-related quality of life (V-RQOL) following type I thyroplasty for unilateral vocal fold paralysis. *J Voice*. 2000;14:378–386.
20. Jones SM, Carding PN, Drinnan MJ. Exploring the relationship between severity of dysphonia and voice-related quality of life. *Clin Otolaryngol*. 2006;31:411–417.
21. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21:576–590.
22. Franco RA, Andrus JG. Aerodynamic and acoustic characteristics of voice before and after adduction arytenopexy and medialization laryngoplasty with GORE-TEX in patients with unilateral vocal fold immobility. *J Voice*. 2009;23:261–267.
23. Sofranko JL, Prosek RA. The effect of experience on classification of voice quality. *J Voice*. 2012;26:299–303.
24. Heman-Ackah YD, Michael DD, Goding GS Jr. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice*. 2002;16:20–27.
25. Awan SN, Roy N. Outcomes measurement in voice disorders: application of an acoustic index of dysphonia severity. *J Speech Lang Hear Res*. 2009;52:482–499.
26. Smits I, Ceuppens P, De Bodt MS. A comparative study of acoustic voice measurements by means of Dr. Speech and Computerized Speech Lab. *J Voice*. 2005;19:187–196.
27. Yiu EM, Murdoch B, Hird K, Lau P. Perception of synthesized voice quality in connected speech by Cantonese speakers. *J Acoust Soc Am*. 2002;112:1091–1101.
28. Moers C, Mobius B, Rosanowski F, Noth E, Eysholdt U, Haderlein T. Vowel- and text-based cepstral analysis of chronic hoarseness. *J Voice*. 2012;26:416–424.
29. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. *J Speech Hear Res*. 1990;33:298–306.
30. Feijoo S, Hernandez C. Short-term stability measures for the evaluation of vocal quality. *J Speech Hear Res*. 1990;33:324–334.
31. Dejonckere PH, Lebacq J. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *ORL J Otorhinolaryngol Relat Spec*. 1996;58:326–332.
32. Wolfe V, Martin D. Acoustic correlates of dysphonia: type and severity. *J Commun Disord*. 1997;30:403–415. quiz 415–406.
33. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice*. 2001;15:529–542.
34. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*. 2005;48:323–335.
35. Choi SH, Zhang Y, Jiang JJ, Bless DM, Welham NV. Nonlinear dynamic-based analysis of severe dysphonia in patients with vocal fold scar and sulcus vocalis. *J Voice*. 2012;26:566–576.
36. Dedivitis RA, Barros APB, Queija DS, Pfuetzenreiter EG Jr, Bohn NP. Achados perceptivo-auditivos e acústicos em pacientes submetidos à laringectomia fronto-lateral; Perceptual and acoustic analysis findings in patients undergone frontolateral laryngectomy. *Rev Bras Cir Cabeça Pescoço*. 2008;37:163–165. [In Brazilian Portuguese].
37. Nawka T. *Die auditive Bewertung heiserer Stimmen nach dem RBH-System*. Stuttgart: Thieme; 1994.
38. Wolfe VI, Steinfatt TM. Prediction of vocal severity within and across voice types. *J Speech Hear Res*. 1987;30:230–240.
39. Halberstam B. Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL J Otorhinolaryngol Relat Spec*. 2004;66:70–73.
40. Boucher VJ. Acoustic correlates of fatigue in laryngeal muscles: findings for a criterion-based prevention of acquired voice pathologies. *J Speech Lang Hear Res*. 2008;51:1161–1170.
41. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *J Voice*. 2004;18:299–304.
42. Pinho SMR, Pontes P. *Músculos Intrínsecos da Laringe e Dinâmica Vocal*. Rio de Janeiro: Revinter; 2008.
43. Moon KR, Chung SM, Park HS, Kim HS. Materials of acoustic analysis: sustained vowel versus sentence. *J Voice*. 2012;26:563–565.